Journal of Physics: Conference Series

PAPER • OPEN ACCESS

A New Robust Estimator to Detect Outliers for Multivariate Data

To cite this article: Sharifah Sakinah Syed Abd Mutalib *et al* 2019 *J. Phys.: Conf. Ser.* **1366** 012104

A New Robust Estimator to Detect Outliers for Multivariate Data

Sharifah Sakinah Syed Abd Mutalib[1,2], Siti Zanariah Satari[1] and Wan Nur Syahidah Wan Yusoff[1]

[1]Centre for Mathematical Sciences, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang, Malaysia

[2]Faculty of Computer, Media and Technology Management, TATI University College, Jalan Panchur, Telok Kalong, 24000 Kemaman, Terengganu, Malaysia.

E-mail: sharifahsakinah84@gmail.com

Abstract. Mahalanobis distance (MD) is a classical method to detect outliers for multivariate data. However, classical mean and covariance matrix in MD suffered from masking and swamping effect if the data contain outliers. Due to this problem, many studies used robust estimator instead of the classical estimator of mean and covariance matrix. In this study, a new robust estimator, namely, Test on Covariance (TOC) is proposed to detect outliers in multivariate data. The performance of TOC is compared with the existing robust estimators which are Fast Minimum Covariance Determinant (FMCD), Minimum Vector Variance (MVV), Covariance Matrix Equality (CME) and Index Set Equality (ISE). The probability that all the planted outliers are successfully detected (*pout*), probability of masking (*pmask*) and probability of swamping (*pswamp*) are computed for each estimator via simulation study. It is found that the TOC is applicable and a promising approach to detect the outliers for multivariate data.

1. Introduction

A multivariate data can be represented by $n \times p$ matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

where x_{ij}, is the value of the *i*th observation for the *j*th variable. n is sample size and p is the number of variables. The population and sample mean vector are represented as $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_p]$ and $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$, where

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad j = 1,2,\ldots,p.$$

The variance-covariance matrix or covariance matrix, $\mathbf{\Sigma}$ is arranged in a $p \times p$ symmetric matrix and is estimated by the matrix $\mathbf{S}$, given by $\mathbf{S} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \Big/ (n-1)$ [1,2], where

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

However, the estimation of sample mean and covariance matrix for multivariate data are affected by outliers, thus have masking and swamping effect [3–7]. Masking effect happens when outliers are identified as inliers (false negative) and swamping effect happens when inliers are identified as outliers (false positive) [3,5,8]. Outliers also can affect multivariate analysis, give incorrect conclusions and makes modelling difficult [3,4,6,9]. Due to these problems, robust methods used to estimate the mean and covariance matrix had been proposed and developed.

Robust estimators are found to be less sensitive and resistant towards outliers compared to classical estimators [3]. Mahalanobis distance (MD) has been used widely to identify outliers in the multivariate data sets and the distance is defined by

$$d_i(\bar{\mathbf{x}}, \mathbf{S}) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})}, \quad i = 1,2,\ldots.,n \tag{1}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean and sample covariance matrix. As for Mahalanobis distance, the sample mean, $\bar{\mathbf{x}}$ and covariance matrix, $\mathbf{S}$ are replaced with the robust estimators of mean and covariance matrix and will yield robust Mahalanobis distance or robust distance [3,9–13].

Various robust estimators had been proposed and developed in the previous studies such as S, M, MM, Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD) and Fast-MCD (FMCD) estimators. Among these robust estimators, FMCD that had been developed by Rousseeuw and Van Driessen (1999) [14] is widely used. FMCD is based on Concentration step (C-step) [10,14] and possess the desirable properties of robust estimators which are affine equivariant, high breakdown point, bounded influence function and has lower computational complexity [10,15–18]. The objective of FMCD is to find *h* observations from the data (subset *h)* whose covariance matrix has the lowest determinant and act as the stopping rule in C-step [10,15,19].

However, the computational complexity of FMCD increases as the number of variables increases [10] and face a singularity problem as the estimator is based on covariance determinant [10]. Therefore, Herwindiati et al. (2007) [10] proposed Minimum Vector Variance (MVV) which used vector variance (VV) as the new stopping rule. MVV can overcome the singularity problem as the computation of MVV is simple, covariance does not need to be positive definite and can be applied to high dimension data sets [10]. MVV also has been proven has the same breakdown point as the MVE and MCD-based methods, robust estimator and has lower computational time than FMCD [10].

Despite the advantages that MVV has possessed, it still has lower computational time as the number of variables increases [20]. This problem has lead Rohayu in 2013 to develop Covariance Matrix Equality (CME) and Index Set Equality (ISE) [21]. CME and ISE are used as the new stopping rule in FMCD algorithm. CME and ISE are able to find the robust mean and covariance matrix [21]. Between these two estimators, ISE is found to have lower computational time compared to FMCD, MVV and CME [21]. Although ISE has superior performance than other estimators and simple to compute, there is no arithmetical computation involved [20].

Basically, CME and ISE are the test of equality between two covariance structures. This motivates us to propose a new stopping rule for C-step which may test the equality of two covariance structures. The new stopping rule is based on Test on Covariance (TOC) method. Final robust estimator of mean and covariance matrix is called TOC. The methods to detect outliers for multivariate data consist two phases. Phase I is to obtain robust mean and covariance matrix and Phase II will use the robust estimator to detect the outliers. The performance of TOC in detecting the outliers for multivariate data will be measured and compared with other robust estimators (FMCD, MVV, CME and ISE).

2. Existing Robust Estimator Algorithms

The algorithm for FMCD is given as follows [21].

Step 1: Select an arbitrarily subset H_{old} containing h different observations, where h is the smallest integer $\geq (n+p+1)/2$, where *p* is the number of variable and *n* is sample size.

Step 2: Compute the mean vector $\overline{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}.

Step 3: Compute $d^2_{H_{old}}(i) = \left(X_i - \overline{X}_{H_{old}}\right)' S^{-1}_{H_{old}} \left(X_i - \overline{X}_{H_{old}}\right)$ for $i = 1,2,\ldots,n$.

Step 4: Sort $d^2_{H_{old}}(i)$ for $i = 1,2,\ldots,n$ in increasing order $d^2_{H_{old}}(\pi(1)) \leq d^2_{H_{old}}(\pi(2)) \leq \ldots \leq d^2_{H_{old}}(\pi(n))$ where π is a permutation on $\{1,2,\ldots,n\}$.

Step 5: Define $H_{new} = \{X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(h)}\}$ and then calculate $\overline{X}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}(i)$ for $i = 1,2,\ldots,n$.

Step 6FMCD: If $\det\left(S_{H_{new}}\right) = 0$, repeat Step 1 – Step 5. Otherwise, if $\det\left(S_{H_{new}}\right) < \det\left(S_{H_{old}}\right)$, let $H_{old} := H_{new}$, $\overline{X}_{H_{old}} := \overline{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop and $\det\left(S_{H_{new}}\right) = \det\left(S_{H_{old}}\right)$ is obtain.

The algorithms for MVV, ISE and CME are the same as FMCD algorithm except for Step 6 [20,21].

Step 6_{MVV}: If $Tr\left(S^2_{H_{new}}\right) = 0$, repeat Step 1 – Step 5. Otherwise, if $Tr\left(S^2_{H_{new}}\right) \neq Tr\left(S^2_{H_{old}}\right)$, let $H_{old} := H_{new}$, $\overline{X}_{H_{old}} := \overline{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$ .Then go to Step 3. Otherwise, the process is stop and $Tr\left(S^2_{H_{new}}\right) = Tr\left(S^2_{H_{old}}\right)$ is obtain.

Step 6_{CME}: If $\sqrt{Tr\left(S_{H_{new}} - S_{H_{old}}\right)^2} \neq 0$, calculate $\overline{X}_{H_{new}}$ and let $H_{old} := H_{new}$, $\overline{X}_{H_{old}} := \overline{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop.

Step 6_{ISE}: If $I_{new} \neq I_{old}$, let $H_{old} := H_{new}$, calculate $\overline{X}_{H_{new}}$ and let $H_{old} := H_{new}$, $\overline{X}_{H_{old}} := \overline{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop.

It has been proven that ISE is effective as FMCD and MVV in finding robust estimator of mean and covariance matrix with lower computational time [20,21]. ISE is only logical comparison of two index sets which are old subset and new subset. The comparison is made to determine the equality of covariance structure for both subsets.

3. A New Robust Estimator Algorithm based on Test on Covariance (TOC)

By adopting the idea of ISE, Test on Covariance (TOC) involving the equality test of variance-covariance structure is proposed. The equality of two covariance structure are tested by using Eq (2) with the hypothesis $H_0 : \Sigma_{old} = \Sigma_{new}$ versus $H_1 : \Sigma_{old} \neq \Sigma_{new}$.

$$u = \upsilon \left[\sum_{i=1}^{p} \left(\lambda_i - \ln \lambda_i\right) - p \right] \tag{2}$$

where $\upsilon = n-1$, $p = 1,2,\ldots,k$ and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\Sigma_{new}\Sigma^{-1}_{old}$. H_0 is rejected if

$u > \chi^2\left[\alpha, \frac{1}{2}p(p+1)\right]$ [22].

The algorithm for TOC is the same as FMCD except a new procedure is proposed for Step 6, Step 6$_{\text{TOC}}$: If H_0 is rejected, calculate $\overline{X}_{H_{new}}$ and let $H_{old} := H_{new}, \overline{X}_{H_{old}} := \overline{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop.

4. Outlier Scenarios and Performance Measures

In this study, only one outlier scenario was investigated. This outliers' scenario is known as mean shift location and given as follows,

$$(1-\varepsilon)N_p(\mu_0, \Sigma_0) + \varepsilon N_p(\lambda\mu_1, \Sigma_1) \tag{3}$$

where $\Sigma_0 = \Sigma_1 = I_p$, $\mu_0 = (0\ 0 \ldots 0)'$ and $\mu_1 = (1\ 1 \ldots 1)'$ is of dimension *p*. Simulation study has been conducted for $n = 200$, $p = 5$, with percentage of outliers, ε ranging from 1% to 25% and the distance of the outliers, $\lambda = 2, 4, 10$ using R language.

Robust mean and covariance matrix from FMCD, MVV, CME, ISE and TOC were used to identify outliers. The steps to identify outliers are given as following,

Step 1: Compute the distance $d^2(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})}$ for $i = 1, 2, \ldots, n$.

Step 2: Use the cut-off value $\sqrt{\chi^2_{p,0.975}}$ in order to detect outliers. If $d(\mathbf{x}_i) > \sqrt{\chi^2_{p,0.975}}$, $\mathbf{x}_i$ is an outlier.

The performance of each robust estimator were measured by *pout* (the probability that all the outliers are successfully detected), *pmask* (the probability that the outliers are falsely detected as inliers) and *pswamp* (the probability of inliers detected as outliers). The best robust estimator will be determining by the highest value of *pout* and the lowest values of *pmask* and *pswamp*. Formulas for these measures are given as follows.

$$pout = \frac{\text{"}success\text{"}}{s} \tag{4}$$

where "*success*" is number of data set that the robust estimators successfully identified all the outliers, and *s* is the total number of simulations.

$$pmask = \frac{\text{"}failure\text{"}}{(out)(s)} \tag{5}$$

where "*failure*" is the number of outliers in all data set that detected as inliers, and *out* is the number of outliers.

$$pswamp = \frac{\text{"}false\text{"}}{(n - out)(s)} \tag{6}$$

where "*false*" is the number of inliers in all data set that detected as outliers, and *n* is sample size.

5. Results and Discussion

The results are shown in Table 1-3 and are illustrated in Figure 1-3. The values of *pout* for distance of outliers, $\lambda = 2$ are decrease as the percentage of outliers increases for all robust estimators. It shows that outliers are easier to detect when the data contain small number of outliers. At 25% of outliers, all robust estimators failed to detect the outliers as the value of *pout* is 0.000. Table 1 show that TOC has the lowest value of *pout* for all percentage of outliers except for 2% and 25%. However, the value of *pout* for distance of outliers, $\lambda = 4$ and $\lambda = 10$ remain 1.000 as the percentage of outliers increases for all robust estimators. As we can see for these cases, different percentage of outliers does not affect the results as the distances get larger. The value of 1.000 for *pout* shows that the estimators successfully identified all the planted outliers.

As can be seen from Table 1, the values of *pmask* of TOC for distance of outliers, $\lambda = 2$ are the highest except for 2% of outliers. This result show that TOC has the highest probability to falsely detected outliers as inliers. However, the value of *pmask* for distance of outliers, $\lambda = 4$ and $\lambda = 10$ shown in Table 2 and 3 remain 0.000 as the percentage of outliers increases for all robust estimators.

TOC shows the lowest value of *pswamp* for distance of outliers $\lambda = 2$ and $\lambda = 4$. This indicates that TOC has the lowest probability to misclassified inliers as outliers. Meanwhile for $\lambda = 10$, TOC has the lowest value of *pswamp* except when the data contain 1%, 2% and 25% of outliers shown in Table 3.

Results in Table 1-3 are illustrated in Figure 1-3 which are easier to examine. Figure 1 shows plot of *pout*. For $\lambda = 2$, the value of *pout* decrease as the percentage of outliers increases and for $\lambda = 4$ and $\lambda = 10$ the value of *pout* remain at 1.000. However, the results for *pmask* in Figure 2 are contrary. For $\lambda = 2$, the value of *pmask* decrease as the percentage of outliers increases and for $\lambda = 4$ and $\lambda = 10$ the value of *pmask* remain at 0.000. Plot of *pswamp* in Figure 3 show similar pattern for all distance of outliers. The value of *pswamp* increases as the percentage of outliers increases.

Summary result for the best robust estimator of each measure is given in Table 4. From Table 4, TOC has problems to detect outliers successfully and misclassified outliers as inliers for distance of outliers, $\lambda = 2$. However, TOC give better performance in *pswamp*. All the results obtained indicate that if outliers are generated close to inliers, it would make the estimators difficult to detect the outliers as the outliers and inliers start to merge.

Table 1. *pout*, *pmask* and *pswamp* values for sample size, $n = 200$, number of variable, $p = 5$ and distance of outliers, $\lambda = 2$.

	pout					***pmask***					***pswamp***				
ε	FMCD	MVV	CME	ISE	TOC	FMCD	MVV	CME	ISE	TOC	FMCD	MVV	CME	ISE	TOC
0.01	0.9892	0.9856	0.9893	0.9880	0.9838	0.0054	0.0072	0.0054	0.0060	0.0081	0.2106	0.2315	0.2267	0.2285	0.2031
0.02	0.9737	0.9734	0.9699	0.9715	0.9726	0.0067	0.0067	0.0076	0.0072	0.0070	0.2374	0.2429	0.2425	0.2418	0.2248
0.05	0.9306	0.9235	0.9235	0.9259	0.8755	0.0071	0.0079	0.0079	0.0076	0.0132	0.2464	0.2455	0.2455	0.2501	0.2248
0.10	0.8438	0.8480	0.8285	0.8285	0.7525	0.0084	0.0082	0.0093	0.0093	0.0141	0.2570	0.2537	0.2533	0.2533	0.2455
0.15	0.7430	0.7040	0.6900	0.7430	0.6160	0.0100	0.0120	0.0120	0.0100	0.0160	0.2700	0.2670	0.2780	0.2700	0.2600
0.20	0.3701	0.3564	0.2648	0.3564	0.1925	0.0245	0.0252	0.0325	0.0252	0.0401	0.2840	0.2882	0.2936	0.2882	0.2834
0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.4775	0.4772	0.4772	0.5512	0.5608	0.3769	0.3808	0.3808	0.3663	0.2835

Table 2. *pout*, *pmask* and *pswamp* values for sample size, $n = 200$, number of variable, $p = 5$ and distance of outliers, $\lambda = 4$.

	pout					***pmask***					***pswamp***				
ε	FMCD	MVV	CME	ISE	TOC	FMCD	MVV	CME	ISE	TOC	FMCD	MVV	CME	ISE	TOC
0.01	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2329	0.2367	0.2328	0.2352	0.2192
0.02	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2495	0.2515	0.2436	0.2495	0.2370
0.05	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2500	0.2526	0.2502	0.2514	0.2380
0.10	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2537	0.2558	0.2558	0.2565	0.2448
0.15	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2566	0.2565	0.2605	0.2565	0.2559
0.20	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2740	0.2740	0.2709	0.2725	0.2690
0.25	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3125	0.3137	0.3137	0.3030	0.3007

Table 3. *pout*, *pmask* and *pswamp* values for sample size, $n = 200$, number of variable, $p = 5$ and distance of outliers, $\lambda = 10$.

	pout					***pmask***					***pswamp***				
ε	FMCD	MVV	CME	ISE	TOC	FMCD	MVV	CME	ISE	TOC	FMCD	MVV	CME	ISE	TOC
0.01	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2117	0.2065	0.2073	0.2082	0.2142
0.02	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2133	0.2142	0.2112	0.2091	0.2197
0.05	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2362	0.2446	0.2414	0.2455	0.2333
0.10	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2478	0.2487	0.2544	0.2544	0.2406
0.15	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2557	0.2670	0.2601	0.2596	0.2476
0.20	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2656	0.2677	0.2681	0.2677	0.2644
0.25	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2729	0.2750	0.2780	0.2718	0.2746

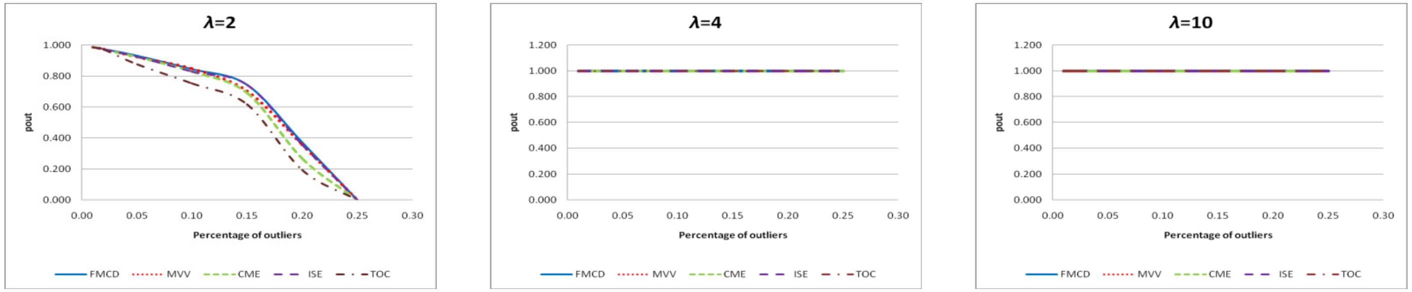

Figure 1. Plot of probability of success (*pout*) versus percentage of outliers with distance of outliers $\lambda = 2, 4$ and 10 .

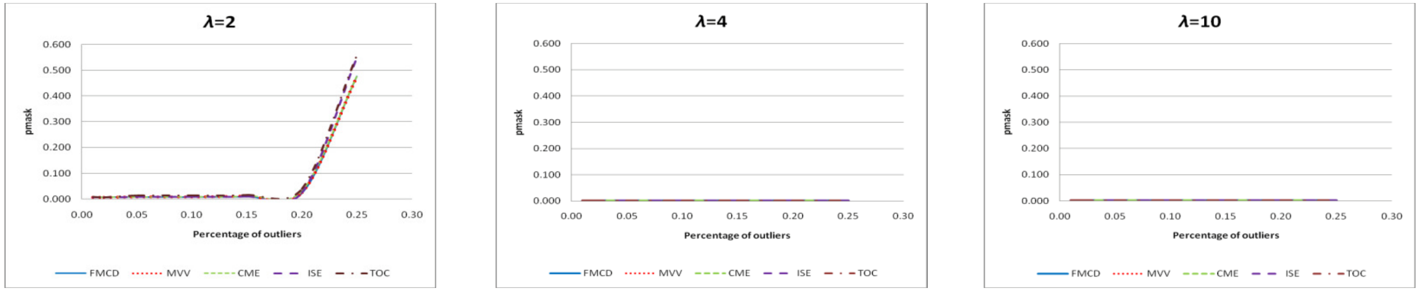

Figure 2. Plot of masking error (*pmask*) versus percentage of outliers with distance of outliers $\lambda = 2, 4$ and 10 .

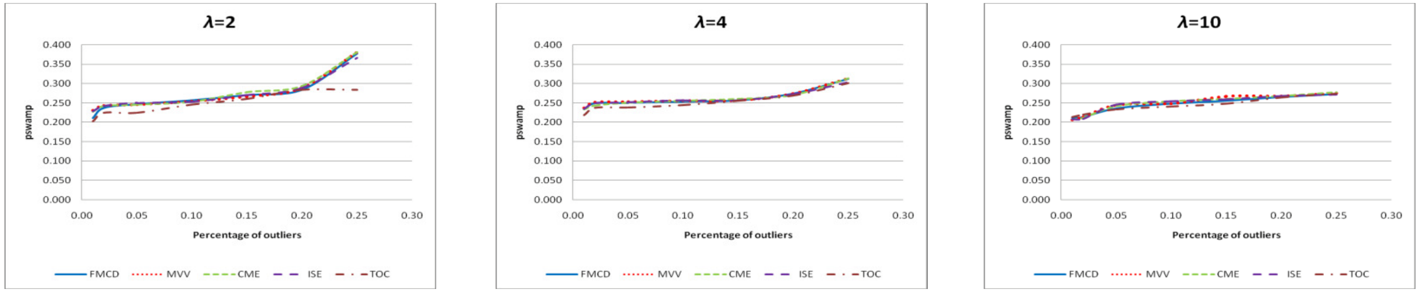

Figure 3. Plot of swamping error (*pswamp*) versus percentage of outliers with distance of outliers $\lambda = 2, 4$ and 10 .

Table 4. The best robust estimators for various percentage of outliers with $\lambda = 2,4,10$

ε	*pout*			*pmask*			*pswamp*		
	$\lambda = 2$	$\lambda = 4$	$\lambda = 10$	$\lambda = 2$	$\lambda = 4$	$\lambda = 10$	$\lambda = 2$	$\lambda = 4$	$\lambda = 10$
0.01	CME	All	All	FMCD, CME	All	All	TOC	TOC	MVV
0.02	FMCD	All	All	FMCD, MVV	All	All	TOC	TOC	ISE
0.05	FMCD	All	All	FMCD	All	All	TOC	TOC	TOC
0.10	MVV	All	All	MVV	All	All	TOC	TOC	TOC
0.15	FMCD, ISE	All	All	FMCD, ISE	All	All	TOC	TOC	TOC
0.20	FMCD	All	All	FMCD	All	All	TOC	TOC	TOC
0.25	-	All	All	MVV, CME	All	All	TOC	TOC	ISE

6. Conclusions

In this study, all the methods shown are the innovation method from FMCD. As can be seen from the algorithms, FMCD used covariance determinant and MVV used vector variance as the stopping rule in C-step. Covariance determinant and vector variance is a scalar representation and do not represent the structure of two covariance matrix. CME and ISE were develop based on the equality of two covariance structure. CME determine the equality of covariance structure based on vector variance. However, in certain cases, two covariance matrix could produce the same value of covariance determinant and vector variance regardless the equality of covariance matrices. Hence, in this study, a new robust estimator which is based on equality test of covariance structure is proposed. It is called TOC.

From the simulation study, results indicate that TOC is considerably good when the distance of outliers increase even though TOC has the highest value of *pmask* and most of the value of *pswamp* are the lowest. This simulation study also shown that the use of TOC is applicable and a promising approach to detect outliers for multivariate data in certain cases. In this study, we only consider moderate sample size and number of variable and also three values for distance of outliers as our preliminary study and results. However, in the future we would like to investigate TOC for different outlier scenarios, distance of outliers, sample sizes and number of variables.

Acknowledgement

The authors would like to thank Terengganu Advanced Technical Institute University College (TATIUC) for the financial support under the grant TATIUC Short Term Research Grant (STG) (9001-1905) and Universiti Malaysia Pahang (UMP).

References

[1] Everitt BS, Dunn G. *Applied multivariate data analysis*. Second Edi. Wiley; 2001.

[2] Johnson RA, Wichern DW. *Applied multivariate statistical analysis*. Fifth Edit. Prentice Hall, Inc.; 2002

[3] Hadi AS, Rahmatullah Imon AHM, Werner M. Detection of outliers. Wiley Interdiscip Rev Comput Stat. 2009;**1**:57–70.

[4] Möller SF, Frese J Von, Bro R. Robust Methods for Multivarite Data Analysis. J Chemom. 2005;**19**:549–63.

[5] Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2011;**1**:73–9

[6] Werner M. Identification of multivariate outliers in large data sets. 2003.

[7] Santos-pereira CM, Pires AM. Detection of outliers in multivariate data: A method based on clustering and robust estimators. 2002.

[8] Filzmoser P, Todorov V. Robust tools for the imperfect world. *Inf Sci* (*Ny*). 2013; **245**:4–20.

[9] Su X, Tsai C-L. Outlier detection. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2011;**1**:261–8.

[10] Herwindiati DE, Djauhari MA, Mashuri M. Robust multivariate outlier labeling. *Commun Stat Simul Comput*. 2007;**36**:1287–94.

[11] Rousseeuw PJ. Least median of squares regression. *J Am Stat Assoc*. 1984;**79(388):**871–80.

[12] Rousseeuw P, Yohai V. Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis.* 1984:256–72.

[13] Rousseeuw P. Multivariate estimation with high breakdown point. *Math Stat Appl.* 1985;283–97.

[14] Rousseeuw PJ, Van Driessen K. A fast algorithm for the Minimum Covariance Determinant Estimator. *Technometrics.* 1999;**41**(3):212–23.

[15] Djauhari MA. A robust estimation of location and scatter. *Malaysian J Math Sci.* 2008;**2**(1):1-24.

[16] Djauhari MA. Highly robust estimation of location and scatter when data sets are of high dimension: An open problem. In: The 3rd International Conference on Mathematics and Statistics (ICoMS-3). Institut Pertanian Bogor, Indonesia; 2008:1–8.

[17] Hubert M, Debruyne M. Minimum covariance determinant. *Wiley Interdiscip Rev Comput Stat*. 2010;**2**(1):36–43.

[18] Hubert M, Debruyne M, Rousseeuw PJ. Minimum Covariance Determinant and Extensions. 2017

[19] Salleh RM, Djauhari MA. Robust start up stage for beltline moulding process variability monitoring using vector variance. *J Fundam Sci*. 2010;**6**(1):67–71.

[20] Salleh RM, Djauhari MA. Robust hotelling's T^2 control charting in spike production process. In: International Seminar on the Application of Science & Mathematics 2011 (ISASM 2011). 2011:18.

[21] Lim HA, Midi H. Diagnostic Robust Generalized Potential based on Index Set Equality (DRGP ISE)) for the identification of high leverage points in linear model. *Comput Stat* [Internet]. 2016; **31**:859–77.

[22] Rencher AC. Methods of multivariate analysis [Internet]. *Wiley-Interscience. John Wiley and Sons Ltd;* 2002.